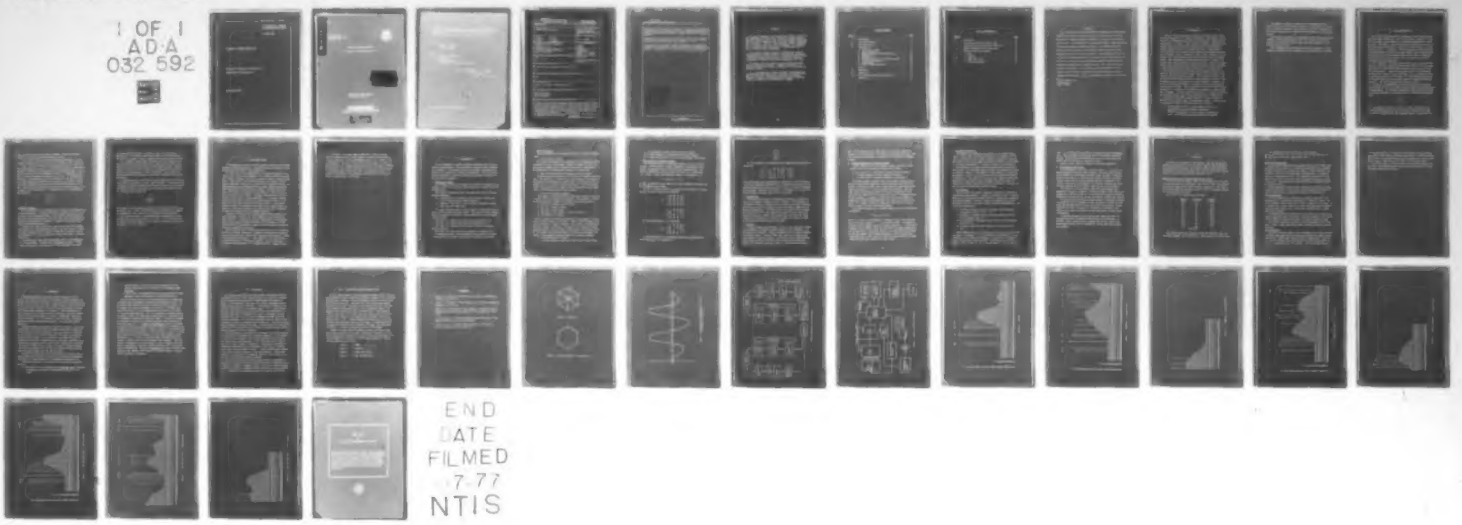
AD-A032 592 PERCEPTION TECHNOLOGY CORP WINCHESTER MASS F/G 5/7
AUTOMATIC SPEAKER ADAPTATION.(U)
SEP 76 H YILMAZ, L FERBER, J SHAO, W PARK F30602-75-C-0227
UNCLASSIFIED RADC-TR-76-273 NL

1 OF 1
AD A
032 592



END
DATE
FILMED
7-77
NTIS

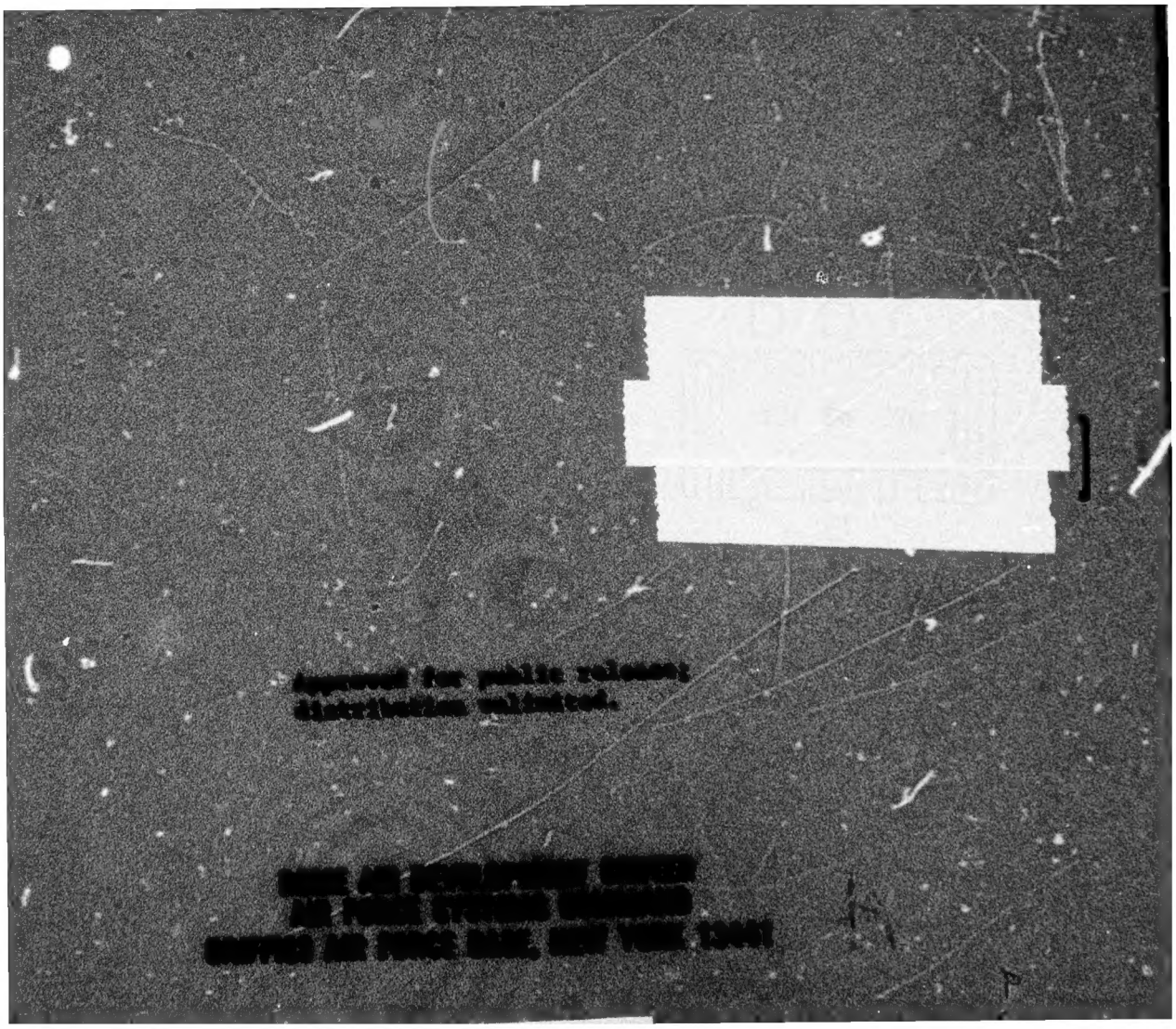
U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

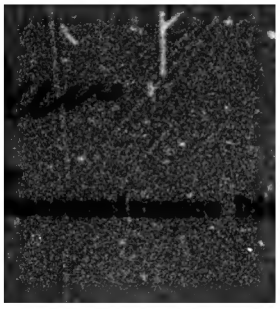
AD-A032 592

AUTOMATIC SPEAKER ADAPTATION

PERCEPTION TECHNOLOGY CORPORATION,
WINCHESTER, MASSACHUSETTS

SEPTEMBER 1976





UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-76-273	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD-A032 592
4. TITLE (and Subtitle) AUTOMATIC SPEAKER ADAPTATION		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report July 1975 - July 1976
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) H. Yilmaz W. Park L. Ferber H. Kellett J. Shao E. Koprucu		8. CONTRACT OR GRANT NUMBER(s) F30602-75-C-0227
9. PERFORMING ORGANIZATION NAME AND ADDRESS Perception Technology Corporation 95 Cross St Winchester MA 01890		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 31011F 70550721
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAP) Griffiss AFB NY 13441		12. REPORT DATE September 1976
		13. NUMBER OF PAGES ~~44~~ 42
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Richard Vonusa (IRAP)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Recognition
Pattern Recognition
Acoustic Phonetics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A speaker-independent speech recognition system was constructed which implements a solution to one of the most difficult and most important problems in speech, that of speaker-to-speaker variability. The system, which recognizes words in naturally spoken, uncontrolled text, is based on a theory of speech perception which is consistent with the linguistic universals of world languages. The representation is invariant under certain adaptive transformations which render the speech speaker-independent. (See reverse)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The problem of speaker-to-speaker variability was solved by reducing the multi-speaker problem to a single-speaker proposition. A single speaker may train the system to recognize a given vocabulary. A subsequent speaker need speak only a predetermined sentence or word sequency to transform the system for operation on his voice.

Performance has been evaluated using constraint-free speech, spoken in natural word sequences. Recognition results for 25 American male speakers are given, indicating an overall recognition accuracy of 97.6%.

It is concluded that the method of speaker transformation has produced marked improvement in the recognition of connected speech, and that the method is applicable to a multiplicity of speech recognition systems for overcoming speaker-to-speaker variability as well as variations due to vocabulary and language.

SUMMARY

A speaker-independent speech recognition system was constructed which implements a solution to one of the most difficult and most important problems in speech, that of speaker-to-speaker variability. The system, which recognizes words in naturally spoken, uncontrolled text, is based on a theory of speech perception which is consistent with the linguistic universals of world languages. The representation is invariant under certain adaptive transformations which render the speech speaker-independent.

The problem of speaker-to-speaker variability was solved by reducing the multi-speaker problem to a single-speaker proposition. A single speaker may train the system to recognize a given vocabulary. A subsequent speaker need speak only a predetermined sentence or word sequence to transform the system for operation on his voice.

Performance has been evaluated using constraint-free speech, spoken in natural word sequences. Recognition results for 25 American male speakers are given, indicating an overall recognition accuracy of 97.6%.

It is concluded that the method of speaker transformation has produced marked improvement in the recognition of connected speech, and that the method is applicable to a multiplicity of speech recognition systems for overcoming speaker-to-speaker variability as well as variations due to vocabulary and language.

TABLE OF CONTENTS

Section		Page
I	Introduction	1
II	The Fourth Function	3
III	Time-Normalization	6
IV	Implementation	8
	1. Extraction of Vowels	8
	2. Signal Processing	9
	3. Representation of Normalized Samples	10
	4. Categorization	11
	5. Data Bank	11
	6. Simulated templates for an Unknown Speaker	12
	7. Preliminary Recognition	13
	8. Final Decision	13
	9. System Hardware Description	14
V	Results	15
VI	Background	18
VII	Conclusions	20
VIII	Some Printouts From the Present System	21
	References	22

LIST OF ILLUSTRATIONS

Figure		Page
1	Vowel Cube	23
2	Vowel Cube Relation to Previous Space.	23
3	Experimentally Determined Fourth Function.	24
4	Speaker-Independent Speech Recognizer.	25
5	Hardware System Block Diagram.	26
6	"...SEVEN...".	27
7	"...FOUR ONE..."	28
8	"...THREE EIGHT EIGHT...".	30
9	"...ONE TWO THREE FOUR..."	32

EVALUATION

One of the most serious problems inherent in Automatic Speech Recognition systems is the extreme variations that occur in the speech signal when the same words are spoken by different individuals. This program addresses this problem area by the implementation of a perceptual theory of speech recognition which minimizes speaker-to-speaker variations. The technique achieves a speaker independent automatic word recognition system by means of speaker transformations based on short time learning. The successful results of this program even though not implemented in real-time show the feasibility of this technique for practical applications involving speaker independent limited vocabulary connected speech recognition systems. With some further improvements and optimizations, it is envisioned that this technique will be invaluable in other areas of Automatic Speech Recognition such as keyword recognition (word spotting), speaker identification, and language classification.

Richard Vonusa

RICHARD VONUSA
Project Engineer

I. INTRODUCTION

The theory of speech perception utilized in this work falls within a more general approach to the problem of perception. This approach is evolutionary in its philosophy and statistical in its methods. The essence of the approach is as follows: First, consider the physical properties of the stimulus energy and its statistical distributions in the environment; second, consider the needs of the organism in terms of individual and social survival. Given suitable neural material and biochemical processes, and given enough time for evolutionary forces to assert themselves, we then postulate that the perceptual devices evolved proceed toward a functional optimum. When supplemented with additional conditions of metabolic and constructional nature, and perhaps some restrictions related to early genetic fixation, the above statements are assumed to provide a suitable foundation to deduce mathematically the overall properties of a perceptual device.

As with all evolutionary processes, the perceptual organization is a matter of compromise and balance between various stimuli in terms of their relevance and statistical distribution. The statistical attitude is here quite basic because perceptual devices are not designed for specific stimuli. Furthermore, we are not interested in specific designs or mechanisms, but rather, in the functional behavior of ensembles of devices under varying distributions of stimuli. We are interested in an optimal functional representation so as to minimize the dependence of perception on speaker vocabulary or language. In other words we are interested in a method of perception and recognition based on the universal characteristics of human speech.

Our general evolutionary adaptive approach to speech perception is described in our previous reports (See Reference Section). Recently significant advances are made in this theory which are as follows:

1. The need and the form of a fourth expansion function is established by detailed perceptual experiments.
2. Studies of variability in rate and manner of speaking have led to a more efficient sampling-normalization procedure.

The addition of these two advances into our understanding led to a better control over the variabilities inherent in speech so that we are now able to recognize continuous speech consisting of a small vocabulary with sufficiently high accuracy to render the machine usable in practical applications.

The system is at present in an experimental stage and has not yet been brought to a real-time operation. This is not a barrier to real-time operation since the method can be turned into machine language within a reasonable length of time.

Many improvements and optimizations are being investigated. We are planning to incorporate these refinements and optimizations in the future.

II. THE FOURTH FUNCTION

Our previous implementations of the perceptual space have been based upon three expansion functions. The first of these is a measure of intensity. The other two make possible a two-dimensional representation after intensity normalization. These two functions are similar to $\sin\phi$ and, $\cos\phi$. The theory allows for additional functions of the form similar to $\sin(n\phi)$ and $\cos(n\phi)$ for $n = 2$ and higher. We have suspected for a long time that, at least one additional function would make possible more accurate distinctions between the continuous speech sounds (vowels, nasals, continuants), but until last year we were unable to find a suitable fourth function satisfying perceptual requirements.

It was suggested by Professor Roman Jakobson during a discussion that linguistic universals emerging from comparative studies of world languages and especially the distinctive feature analysis seem to imply an 8-vowel cubic representation. The vowel cube so conceived may be considered as existing inside a spherical perceptual space. This is a generalization of our previous speech circle into a sphere. The spherical perceptual space is obtained by utilizing four expansion functions (three plus normalized intensity). The resulting vowel cube is shown in Figure 1.

In the cubic representation, the eight basic sounds are associated with the eight vertices of the cube, and pairs appearing on opposite vertices are complementary. The following complementary pairs have been previously verified experimentally in connection with our two-dimensional representation.

u ↔ ɛ
o ↔ i
a ↔ ʉ

The phoneme ö (similar to the vowel in bird) is predicted by the distinctive feature theory as one of the other two sounds for the vowel cube, but the identity of the last remaining sound is not possible to pre-

dict by the distinctive features theory uniquely. We have undertaken to identify this sound and have performed various tests.

If the cube is projected onto two dimensions perpendicular to a line through two opposite vertices, the result is a hexagon as shown in Figure 2.

If ö is at the top (closest to the reader's eye) then ü, o, and ɛ are on a higher level than u, a, and i. This feature is used in a computerized method of identifying the fourth function of the perceptual space. It was then clear that the missing complementary (namely, the complementary to ö) must be a sound simultaneously similar to u, a, and i. By repeated experimentation, a complementary sound to ö was found and it was verified by inverse filter listening experiments and by its ability to produce the theoretically predicted fourth function when used with other phonemes of the cube. This sound that passed both of these tests is similar to nasal ŋ (as in king), but in sustained form. It completes the set of 4 complementary pairs which are

u ↔ ɛ
o ↔ i
a ↔ u
ö ↔ ŋ

Listening Experiments

An ö was spoken into the microphone and its spectrum was obtained on the PTC spectrum analyzer. The spectral display was carefully traced in pen on the face of the oscilloscope display. A flat spectrum (narrow pulse) was next fed to the spectrum analyzer, and filter gains were adjusted until the spectrum exactly matched the original ö. Listeners verified that perceptually, the resulting sound was ö.

Listeners were then asked to first listen to the ö until they became fully adapted to it, then immediately switch to the flat spectrum. Through perceptual adaptation the flat spectrum is expected to assume the form of the complement of ö. Indeed listeners most often heard nasals, ŋ and sometimes m or n.

In a similar way, the reverse relationship (that ö is the complement of m, ŋ) was verified. In almost every case listeners heard ö. In these contrast experiments, it was found to be of some importance that subjects

were convinced of the identity of the first sound before switching to the flat spectrum. The resulting sound always seemed to be the complement of what the subject thought he heard, rather than what was physically presented.

As an independent experimental verification of these concepts, correlations resulting from a word recognizer were interpreted as cosines of the angles of vectors representing them in a four dimensional space. Then the recovered distances are used to construct a three-dimensional figure which turned out to be approximately a cube, as predicted. These results are being implemented into the program of word speech recognition for practical applications.

A computer program was written for directly graphing the fourth function as obtained from one person or a small number of people. The method consists of summing the spectra of sounds above the center plane of the vowel cube and summing those below the center plane, then obtaining the difference of the two. Sounds actually spoken were

$$\begin{array}{c} \ddot{o} \\ \hline o,\ \varepsilon,\ \ddot{u} \\ \hline u,\ a,\ i \\ \eta \end{array}$$

Four examples of each phoneme were spoken by each speaker covering the range of normal pitches. Figure 3 gives the result as averaged over ten speakers. The resemblance of this curve to a sin 2ϕ function is quite apparent.

In the implementation of a four-dimensional representation one could therefore chose, as primary vowels, u, a, i and ö. If the vocabulary does not contain ö one could replace ö with ε without sacrificing anything in the representation of that vocabulary.

III. TIME-NORMALIZATION

Time normalization is part of the more general problem of how to sample speech for efficient recognition. The general problem of sampling has many aspects only one of which is time-normalization.

The basic idea of time normalization is to sample the speech so as to render it more or less time independent. It is usually achieved in a crude way by taking the samples only after a significant amount of spectral change occurs. This, however, is not sufficient because the intensity, rate of change of intensity, voicing etc. are part of the recognition criteria. The time-normalization was therefore improved by adding additional parameters which represent the influence of these variables. Intensity and rate of change of intensity are controlled by two parameters. Voicing is used to label voiced and unvoiced samples. Additional improvements are made in the fricative and gap areas to reduce oversampling. In the present sampling procedure we have utilized correlation, peak normalized intensity, voice-unvoice and the channel characteristics during silences. The improved sampling procedure so obtained has been tested and is working fairly satisfactorily. The computer printouts of recognized words given elsewhere in the report are produced under this sampling procedure.

We point out, however, that even this improved sampling method is not fully satisfactory, because it is still dependent on how saturated the spoken words are. In particular, distinctly pronounced or saturated words result in more samples than the ones which are not saturated. The main problem caused by such mismatch of samples is that they get out of step and occasional recognition errors occur.

There are various ways of overcoming this problem. One is to subdivide the word into smaller, phoneme size, components and prevent the matching as a whole from getting out of step by first processing the components. Another is to provide word alternatives. A third would be to take into account the saturation explicitly in the course of sampling. This is equivalent to introducing another parameter to the sampling procedure. Roughly speaking

this is analogous to an image sharpening operation. We have devised a method of renormalizing the syllabic segments to achieve the equivalent of an image sharpening or contrast enhancement process. This, however, is not yet implemented. When all of these are done the procedure is, however, no longer equivalent to a standard time-normalization or time-warping technique. Our experience to date seems to show that one of the most crucial component of a continuous speech recognition system is the sampling procedure. We believe we have achieved a reasonably good sampling process and expect to improve it significantly in the future.

IV. IMPLEMENTATION

We have implemented a speaker-independent connected speech recognizer based on a generalization of the techniques developed and tested under previous contracts. The techniques include the perceptual representation of vowels and its application to speaker transformations. We have also developed and implemented a time-normalized sampling procedure and applied it to connected speech recognition.

1. Extraction of Vowels

We have built the necessary hardware and devised the necessary software programs to extract a set of four suitable vowels from a connected utterance of common words.

The concept of extracting vowels from common words has the following practical advantages:

a) A speaker need not be trained to say the vowels, instead, only a set of common words are required from him.

b) A speaker is more likely to give consistent utterances in common words than in isolated vowels.

The present software requires a prescribed set of common words which we chose to be the sequence "one three seven". Upon receipt of this utterance, which can be spoken in a discrete or connected manner, the system proceeds to extract vowels.

The vowel "U" is taken at the onset of the voiced portion of 1.

The vowel "A" is taken at the dominant voiced portion of 1.

The vowel "I" is extracted from the region following the dominant voiced portion of 3.

The vowel "E" is extracted from the dominant voiced portion of 7.

The extracted vowels, each represented by their spectrum, are stored in the form of filter outputs, which contain 16 numbers for each vowel.

2. Signal Processing

The system processes all incoming utterances in the manner described below:

A. Fixed interval sampling.

The system is designed to have the capability of processing a signal duration of 2.25 sec. at each entry. During the allowed 2.25 sec. the system takes a reading of the 16 analog filters every 10 milliseconds, giving maximum of 225 fixed-interval samples for each utterance.

B. Noise level and detection of beginning and end of signal.

Silence portions are monitored for the purpose of determining the noise level. The noise level is taken to be the time average of noise energy in the channel during the absence of speech sounds. The threshold level for signal is set at 1.3 times the detected noise level. The first and last signals having energy greater than the threshold level are marked, respectively, as the beginning and end of an utterance.

C. Selection of Normalized Samples.

Final samples are selected from the fixed-interval samples, starting at 8 samples before the beginning of an utterance. The selection is based on the time-normalization procedure and includes the following factors:

a) The amount of change in spectral shape
b) The rate of change in energy
c) The level of signal energy
d) The nature of the signal i.e., voiced or unvoiced
e) The duration of the signal

All these variables are monitored sequentially from one sample to the next and the combined changes are calculated. When the combined changes exceed a preset criterion, a sample is selected. This process is repeated until the whole utterance is exhausted. In general, in an utterance of three connected numerals, the total number of such samples varies between 30 and 40. The numbers are nearly independent of time but vary with habit and dialect. This method of selection performs the following functions:

a) It time-normalizes the samples
b) It treats transition regions and steady vowels on equal footing

c) It distinguishes between voiced and unvoiced signals

d) It places emphasis on the sequential ordering of samples

e) It takes into account the rate of time-development of the signal.

3. Representation of Normalized Samples

Each normalized sample is originally obtained as a set of 16 filter readings. For the purpose of speaker-independent transformation, these samples are expanded in terms of the four vowels extracted above. The resulting representation for the sample then consists of 4 coefficients α, β, γ, and λ. Each of the normalized samples is represented as in

$$P = \alpha U + \beta A + \gamma I + \lambda E$$

In order to represent a given sample of 16 filter readings by the above form the following steps are taken:

a) The 4 x 4 symmetric Matrix containing the correlations, XY, between any pair of the base functions is calculated

$$M = \begin{vmatrix} UU & AU & IU & EU \\ UA & AA & UA & EA \\ UI & AI & II & EI \\ UE & AE & IE & EE \end{vmatrix}$$

$$= \begin{vmatrix} 1 & AU & IU & EU \\ UA & 1 & IA & EA \\ UI & AI & 1 & EI \\ UE & AE & IE & 1 \end{vmatrix}$$

b) The inverse Matrix M^{-1} is calculated

$$M^{-1} = \begin{vmatrix} 1 & AU & IU & EU \\ UA & 1 & IA & EA \\ UI & AI & 1 & EI \\ UE & AE & IE & 1 \end{vmatrix}^{-1}$$

c) The column Matrix of correlations between the normalized sample, P, and the vowels above are calculated as:

$$\begin{vmatrix} PU \\ PA \\ PI \\ PE \end{vmatrix}$$

d) The coefficients of expansion for the normalized sample can then be obtained as

$$\begin{vmatrix} \alpha \\ \beta \\ \gamma \\ \lambda \end{vmatrix} = \begin{vmatrix} 1 & AU & IU & EU \\ UA & 1 & IE & EA \\ UI & AI & 1 & EI \\ UE & AE & IE & 1 \end{vmatrix}^{-1} \cdot \begin{vmatrix} PU \\ PA \\ PI \\ PE \end{vmatrix}$$

It can be shown that this representation is equivalent to first constructing a set of four orthogonal functions and then representing the sample by these orthogonal functions as long as the choice of vowels is linearly independent and perceptually consistent with section II. For example U, A, I and aU + bI cannot be chosen as primary vowels, as M^{-1} will vanish.

4. Categorization

Categorization is aimed at circumventing the problem of speaker variations, in the manner of speaking, accent or dialect. The differences in vocal characteristics are the parts that are removed by the above transformations which do not affect non-phonetic variabilities such as dialectual and habitual idiosyncracies. The categorization is a process by which a speaker can be placed in one of a few categories according to accent, dialect and habitual differences. Whether a new speaker falls into one of the chosen categories is determined by the degree of closeness with which his common word characteristics match those in the category.

5. Data Bank

For each category described above, we store in our data bank the expansion coefficients of normalized samples belonging to the vocabulary to be processed by the recognizer. These are gathered from the speakers belonging to the same category. In general there are a great deal of similarities among speakers in the same category. In those cases where large deviations occur alternative forms of the same word resulting from idiosyncracies, are stored. The systematic gathering of data is a time-consuming and tedious

task. We have gathered in our data bank up to this time a total of 14 categories for the ten digits. Of these only 8 categories were used. Without the speaker transformations each speaker would constitute a different category.

6. Simulated templates for an unknown speaker

The basic idea here is that an unknown speaker's templates for the vocabulary words can be simulated by knowing the following:

a) His vowels U, A, I, and E (or ô if occurs in the vocabulary)
b) A category closest to his own vowel characteristics
c) The coefficients of expansion belonging to the category.

The unknown speaker's primary vowels are obtained by requiring him to say a set of prescribed common words such as "one three seven" into the system. A better set would be "she too must learn" but time-window limitations did not permit its use for all speakers consistently. Thus the common words are judiciously chosen but they are not necessarily to be taken from the vocabulary words as long as they contain the necessary vowels for the vocabulary. A category is selected for the unknown speaker by comparing his common word characteristics with those in the existing categories. The category in which the common word characteristics are most alike is taken to be that of the unknown speaker.

The normalized samples for the unknown speaker are then simulated using the equation

$$P = \alpha U + \beta A + \gamma I + \lambda E$$

where α, β, γ and λ are coefficients of expansion stored in the data bank under the category closest to that of the unknown speaker. The sequential ordering of these computed samples are strictly adhered to. In this way the templates of the vocabulary words are created for the unknown speaker without his specifically training the system. Since the templates are created by using the speaker's own category functions as they occur in his own common words, only a small amount of variability remains between his actual templates and the ones created by the above process.

7. Preliminary Recognition

The capability of simulating the templates for the unknown speaker renders a single speaker recognizer conducive to multi-speaker use. The recognition algorithm in this system is therefore designed for achieving high accuracy in the single speaker case and its extension to multi-speaker use is through transformation to categories existing in the data-bank.

The utterances of an unknown speaker undergoes signal processing as stated earlier. The resulting normalized samples are compared with those in the simulated templates. Time-sequence of these samples in a given block and sequence of blocks in a given word are strictly maintained throughout the comparisons. The section of an utterance that compares favorably with certain words, (the figure of merit exceeding a prescribed value), is assumed as one of these alternative words but no decision is yet made. We call this stage the preliminary recognition stage.

8. Final Decision

At the end of the preliminary recognition stage only a few possible outcomes await final decision. In fact if the figure of merit is set high enough most of the words are already reduced to a single choice, hence they are already recognized. There are, however, few remaining cases where further decisions are to be made to resolve conflicts and ambiguities. These are of the following type:

a) A short word matching with part of another word and causing a "phantom", such as 3 → 38, 7 → 71.
b) The number of samples being too large and reducing the score, due to mismatch, such as 7 → ?
c) A long template "swallowing" a short word due to fast speaking, such as 38 → 3.
d) The parts joining two words triggering a third one, such as 34 → 0.

These ambiguous cases are resolved by what we call, the "final editing" procedures. For example 3 → 38 is corrected for the "phantom" 3 by a program which will not allow an 8 followed by a 3 unless that 8 is higher in figure of merit than a preset threshold. This threshold is such that it allows a "true" 8, hence 38 → 38 is secured whereas 3 → 38 is corrected

into 3 + 3. Similar editing procedures are applied for the other ambiguous cases. These procedures are explicitly given in the overall recognition program. Those confusions and ambiguities we could not overcome at the present time are considered as recognition errors.

Figure 4 shows a schematic diagram of the system configuration.

9. System Hardware Description

A block diagram of the system is shown in Figure 5. It consists of a Digital Equipment Corporation PDP-8E computer with twenty eight thousand words of memory and an Extended Arithmetic Element Type KE8-E, 512K fixed head disk, dectape drive, a tektronics CRT graphics terminal, speech processing circuits and the interface between them and the computer. The system is activated at the beginning of an utterance, processes the string of words and prints out their identity after the end of the utterance.

The signal from the microphone is amplified, high-frequency pre-emphasized, and passed through the 24 dB/octave bandpass filter with cut-off frequencies at 250 Hz and 5300 Hz. From the resulting signal, two types of information are extracted, the spectral distribution, and auxiliary features.

The spectral distribution of the speech signal is determined by passing it through the bank of 16 bandpass filters, the outputs of which are rectified, smoothed, sampled every 10 msec, and stored in the computer. Linear combinations of these 16 channels are calculated and tabulated in the computer to form the data points in the four-dimensional, frequency domain representation.

A specially designed circuit involving audio compression and zero-crossing information distinguishes between noise and an utterance. It provides a binary waveform which is sampled every 10 msec, stored, and, under program control, used to determine the voicing state and the end of the utterance.

V. RESULTS

The results are presented in four sections. Each section describes evaluation tests conducted to test a specific stage in the development of the speech recognition system. The first two sections cover preliminary evaluation of a partially completed system, the last two sections show performance of the final version of the recognition system.

Recognition Results for a Selected "Hard-Set" of Digits

In order to test the approach under a severe condition a set of difficult connected digits was established. The strings of digits were selected based on past experience with other methods of recognition. The strings were chosen because they presented problems in previous recognition schemes due to coarticulation and stress.

The results for 10 speakers are shown below:

Speaker	# of Digits	% Correct
B.P.	48	95.8
R.V.	42	95.3
A.K.	75	97.4
L.F.	53	100.0
W.B.	75	98.7
H.K.	54	100.0
N.J.	69	98.6
W.S.	75	97.4
R.W.	39	95.0
F.D.	30	96.7
Total	561	

The results show overall accuracy of 97.4% for individual digits and 93% correct sequences for the "hard set". There were 187 sets and 13 errors.

The sequences used for the "hard set" were as follows:
118,111, 311, 318, 418, 411, 711, 718, 911, 918, 831, 838, 839, 841, 848, 849, 859, 088, 188, 288, 388, 488, 788, 888, 988.

Preliminary Demonstration

The preliminary demonstration was based on twelve sets of random digits recorded by twelve speakers, two of which were from RADC. The recordings consisted of twenty sets of three digit strings. Of the 720 digits, 1 was an omission error and 4 were extraneous errors (phantoms). The overall accuracy of recognition was 96.1%. The extraneous errors as well as the omission errors are due to the fact that the number of digits in a string is not known. The program scans through the data and any number of digits is likely to come up.

Close examination of these results indicated that the phonetic editing programs were deficient. The phonetic editing programs were rewritten and a set of editing rules was implemented before the final evaluation.

Final Demonstration

The final demonstration consisted of a technical session and a live demonstration of recognition of English digits in connected strings. Each of seven speakers read a random list of digits in groups of four, three, two or one digits per string. Most of the total of 282 digits were in groups of three digits per string (70%). The other 30% consisted mostly of single or double digit strings.

The demonstration was conducted live so that when an error occurred the speaker repeated the same string again. Using this procedure it was possible to test whether the system can be used as a practical data entry system.

The overall recognition accuracy for the seven speakers was 97.5% per digit. After a single repetition of each of the error-strings the accuracy was 99.3%.

Final Evaluation

In order to obtain a higher level of confidence in the results obtained during the final demonstration the system was retested for twenty five (25) male speakers. Each speaker recorded a list of random digits. There were two sets of recordings one set was recorded at Perception Technology Corporation and contained sixteen (16) speakers. The PTC recording consisted of

150 digits per speaker, 20 strings of triple digits, 25 strings of double digits and 40 single digits recorded in random fashion. The RADC recordings were recorded in the same manner for nine (9) speakers except for five additional strings of three digits for a total of 175 digits per person.

The results indicate that only three of the twenty five speakers were below 95%. The average recognition score including all sources of errors and rejections was 97.6% per digit.

VI. BACKGROUND

Perception Technology Corporation has been working for the last several years toward the solution of the problem of speech perception and recognition under various conditions of channel distortion and speaker variability. Our past work in the area of speech perception could be described as an effort toward invariant extraction of relevant parameters of human speech under various conditions of variability. These variabilities are partly external, such as channel noise and distortion, and partly internal such as intra- and inter-speaker variability, interphonemic interaction, accent, and dialect. In its most general form the problem is formidable, especially since the human perceptual process is not completely known or understood.

Perception Technology Corporation pursued the solution of this problem through what we believe an effective combination of theoretical and experimental research into speech perception. What we have done is essentially the application of the "scientific method" so well known to be operative in positive sciences, namely, to first formulate a plausible theory of the phenomenon based on what is already known, and then test this theory by new experiments it suggests to find out more, and to determine its limitations. As new facts are uncovered one is then in a position to improve, modify or alter the original theory until all the facts, previously known and newly uncovered, may be summarized by the improved theory.

Perception Technology Corporation went through this process, and by so doing produced what we believe a theoretically coherent and experimentally viable theory of speech perception at the phonetic and phonemic level.

Three main conclusions of the theory and the supporting experimental data are:

a) That there exists a multidimensional <u>perceptional space</u>, independent of language or of speaker, in which speech sounds can be represented.

b) That this space is not in one-to-one correspondence with the physical signal-space but is defined up to some adaptive transformations.

c) That at least in the phonemic level speech sounds tend to be categorized to warrant the application of statistical methods.

We have shown that this work is directly applicable to the objectives of implementing an operational continuous speech recognition system. The perceptual space delineates the extent to which the physical signal is to be expanded into linearly independent base functions. The present data shows that the number of such independent functions may not be more than 4-5 for speech intelligibility. For full naturalness and speaker identity the number is larger but probably not larger than 10-15. The simplest case of 3 independent functions was extensively studied. The case of 4 independent functions is found to be necessary for higher accuracy in intelligibility and recognition. The adaptive transformations imply the invariance of perceptually relevant parameters under the conditions of variability such as channel distortion and speaker-to-speaker variations. The adaptive transformations have the dimensionality of the space itself, namely, if 4 independent functions are chosen as adequate for a given purpose then the transformations are 4 x 4 matrices. The intra-speaker transformations may be viewed as a special case of the inter-speaker transformations. The inter-phonemic transformations are more complex in nature although in some sense they may be regarded as short-time (context dependent) limit of the intra-speaker transformations. Finally the categorical perception of speech sounds imply the large tendency of discreteness of perceptual response to speech sounds, especially to consonants.

VII. CONCLUSIONS

The present system demonstrates that speaker transformations reducing the variabilities due to vocal characteristics can be achieved by using perceptual expansion functions and their transformations from speaker to speaker. The key element is the choice of a linearly independent and perceptually significant set of functions. If the number of functions is too small the representation is not accurate enough. If it is larger than perceptually required for speech intelligibility then the representation is too detailed, which makes the categories very large. Furthermore, too many functions often tend to be not linearly independent and make the matrix inversion meaningless. Four linearly independent functions seem to be both perceptually required and computationally trouble free. Accuracies achieved, although somewhat below human performance, are high enough to warrant further elaboration of the system to produce a practical, highly accurate continuous speech recognizer for a small vocabulary.

Many improvements, optimizations and refinements that we can incorporate were not possible to implement due to program limitations. We are planning to enlarge the time window so that a four word preamble such as "She too must learn" can be used consistently. This would also allow the entry of four or five digit strings for recognition. The parameters introduced for recognition purposes are not fully optimized. They were set to certain values after a limited number of trials. To really optimize these parameters we must first bring the machine to a real-time operation. The real-time operation is also essential for gathering statistics and of course for eventual use of the machine for practical purposes.

The conclusion seems to be that the present method is applicable to a multiplicity of speech recognition systems and reduces the variabilities due to speaker, vocabulary or language as far as the vocalic (acoustic-phonetic) aspects are concerned. Alternative pronounciations due to accent and dialect are not covered by the transformations when the variants are sufficiently different.

VIII. SOME PRINTOUTS FROM THE PRESENT SYSTEM

To give actual examples of how the present continuous speech recognition system operates we have included some printouts. Examples of one, two, three and four digits are shown. The bars with digits on the top show the likelihood scores of each digit. The digits with stars are the ones actually selected by the system. The numbers at the far end of the bars correspond to the numbers of normalized samples in the corresponding digits. The numbers at the bottom of the bars are absolute scores. The recognition is made on the basis of bar lengths which are relative scores. The use of relative score (relative to the highest in the utterance) minimizes the omission errors. The (-) signs correspond to unvoiced areas. The system evaluates the channel noise before each utterance. The beginning and end of sampling are governed by this evaluation. No samples are taken unless the energy is 1.3 times the average noise level of the channel.

Figures 6 through 9 show some examples of one, two, three and four digit strings processed by the system through final recognition. The four digit strings must be pronounced quite fast as the total time-window is only 225 miliseconds, part of which is used for channel evaluation. The illustrated utterances are as follows:

Figure 6 - "...SEVEN....."

Figure 7 - "...FOUR ONE....."

Figure 8 - "...THREE EIGHT EIGHT...."

Figure 9 - "...ONE TWO THREE FOUR..."

REFERENCES

1. Yilmaz, H., "A Theory of Speech Perception", Bulletin of Mathematical Biophysics, vol. 29, 1967.

2. Yilmaz, H., "A Theory of Speech Perception II", Bulletin of Mathematical Biophysics, vol. 30, 1968.

3. Yilmaz, et.al., "A Real Time, Small-Vocabulary, Connected-Word Speech Recognition System, Final Report RADC-TR-72-281, November 1972 (AD75 3176).

4. Yilmaz, H., et.al., "Perceptual Continuous Speech Recognition", Final Report RADC-TR-74-180, July 1974 (AD783899).

5. Yilmaz, H., et. al., "Speech Perception Research", Final Report, Contract No. DAAB03-72-C-0407, March 1973.

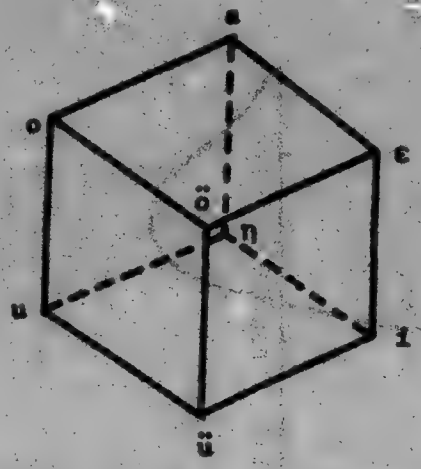


Figure 1. Vowel Cube

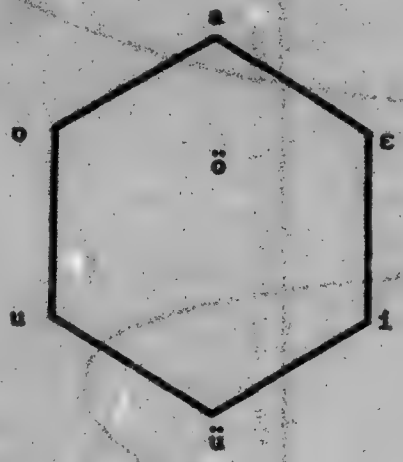


Figure 2. Vowel Cube Relation to Previous Space

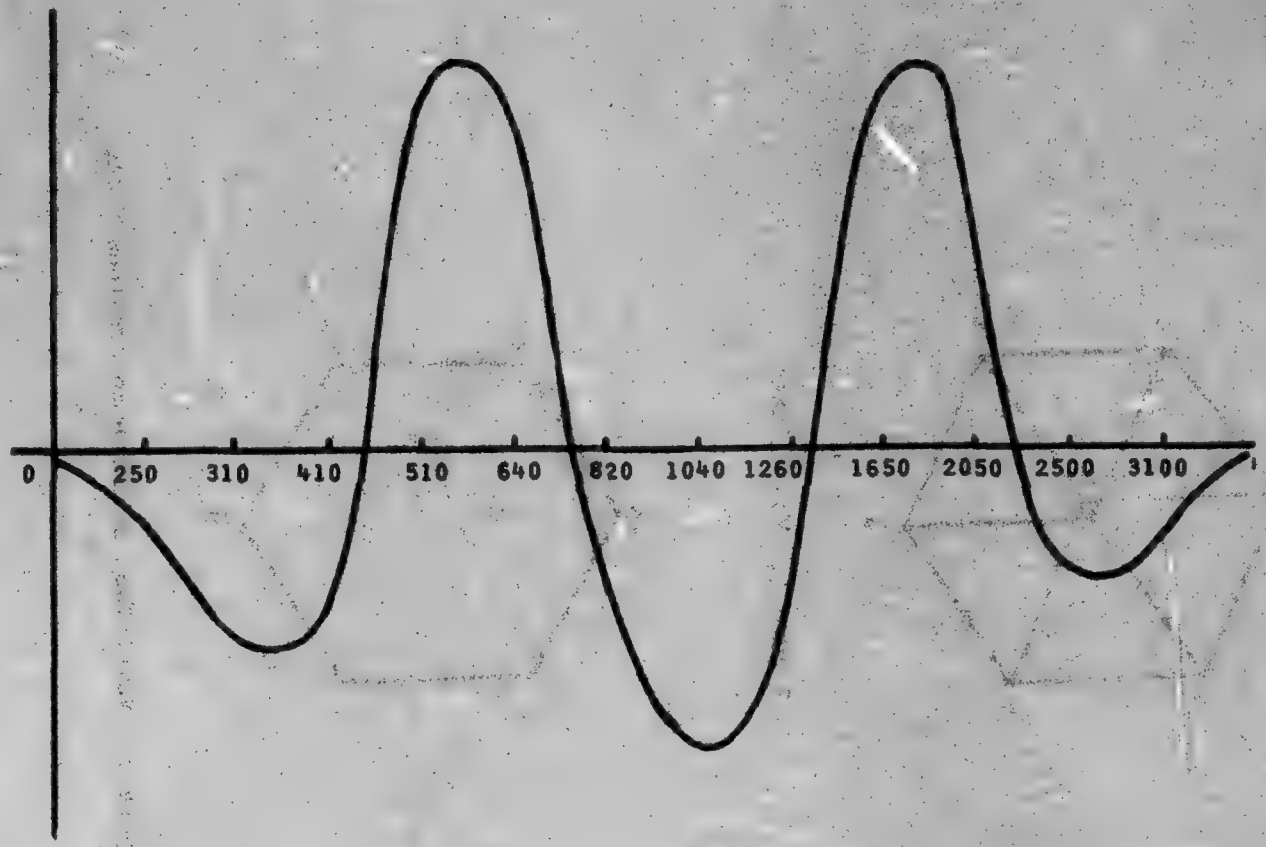


Figure 3. Experimentally Determined Fourth Function
(average for ten speakers)

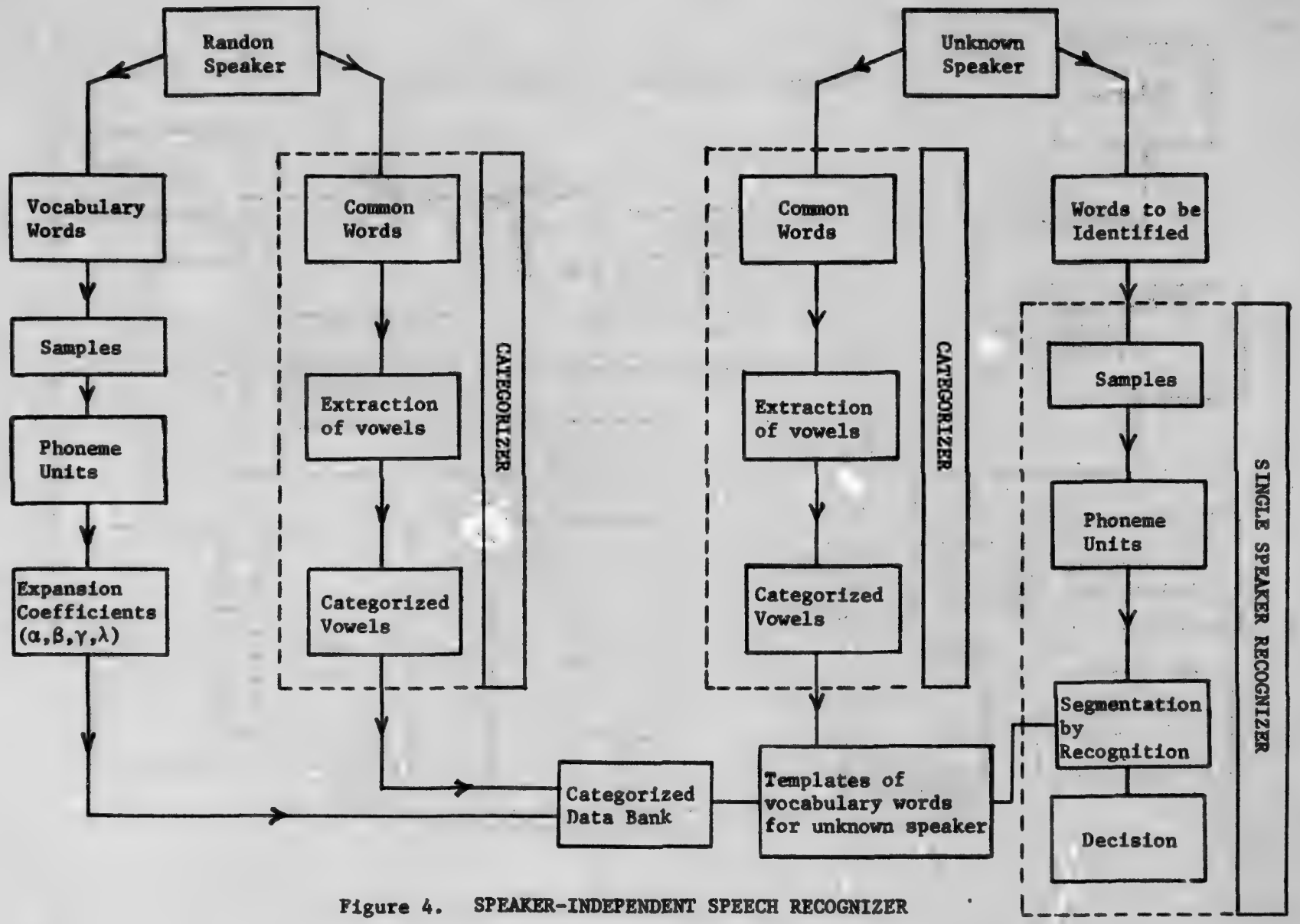


Figure 4. SPEAKER-INDEPENDENT SPEECH RECOGNIZER

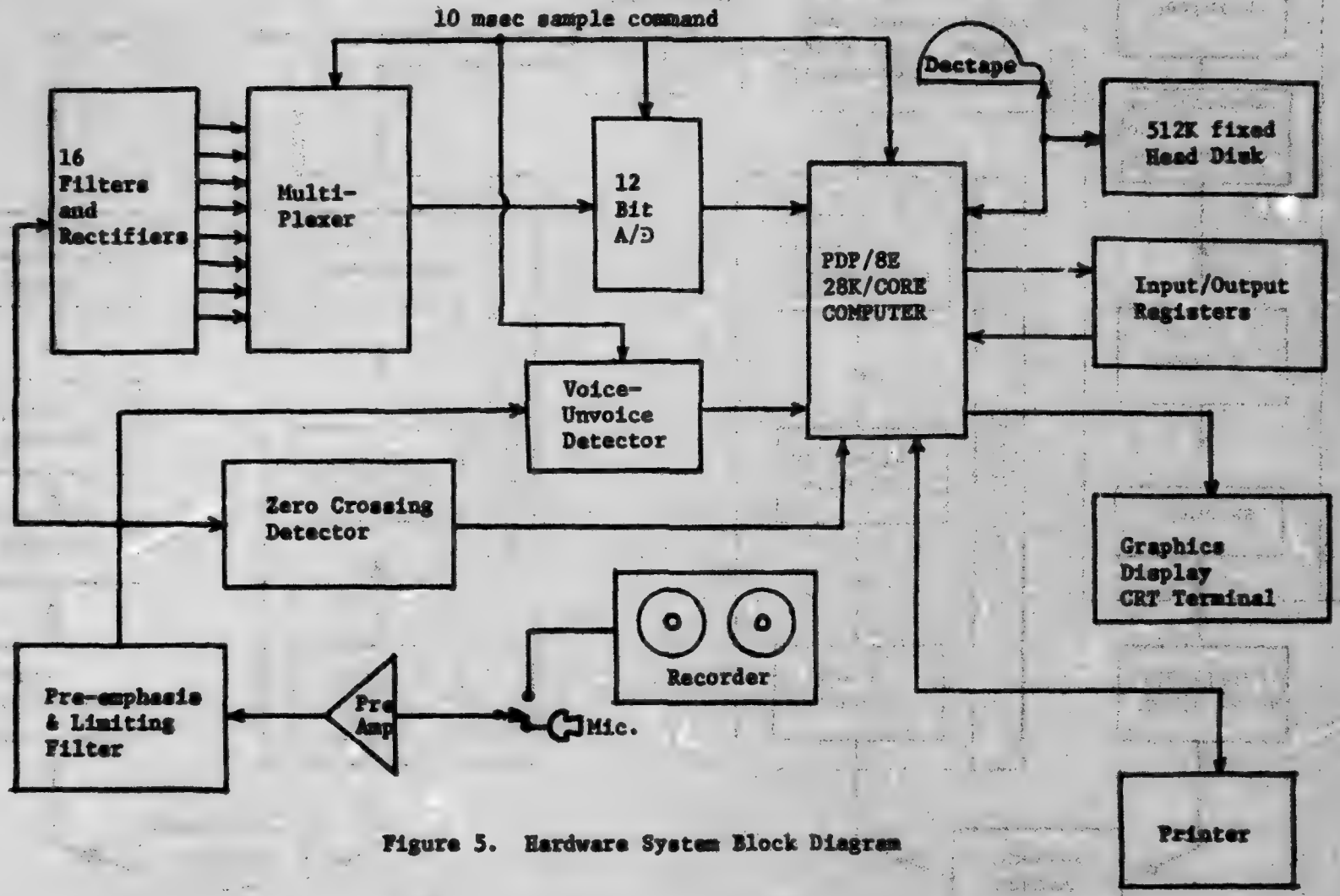


Figure 5. Hardware System Block Diagram

Figure 6: "...SEVEN..."

```
TIME INTERVAL SAMPLES: FIRST= 21 LAST=111 NOISE LEVEL= 216.

THE WORDS RECOGNIZED ARE 4 1

21 -1  0 0.00 =*
22 -1  0 0.00 =*
23 -1  0 0.00 =*
24 -1  0 0.00 =*
25 -1  0 0.00 =*
26 -1  0 0.00 =*
27 -1  1 2.93 =* ---------------------------------4                10
28 -1  0 0.00 =*
29 -1  2 3.25 =* ----------------------------------4                10
30 -1  0 0.00 =*
31 -1  0 0.00 =*
32 -1  0 0.00 =*
33 -1  0 0.00 =*
34 -1  3 3.55 =* ------------------------------------------4*       10
35 -1  0 0.00 =*
36 -1  0 0.00 =*
37 -1  0 0.00 =*
38 -1  4 3.32 =* ----------------------------------5                10
39  1  5 3.37 =====* ---------------------------------1                9
40  1  0 0.00 ========*
41  1  0 0.00 ===========*
42  1  6 3.26 =============* -------------------------1                9
43  1  0 0.00 ===============*
44  1  0 0.00 ===============*
45  1  0 0.00 =================*
46  1  7 3.12 ==================* -------------------1                9
47  1  0 0.00 ======================*
48  1  0 0.00 =======================*
49  1  8 2.92 ========================* -----------9                 8
50  1  0 0.00 ========================*
51  1  0 0.00 ==========================*
52  1  0 0.00 ==========================*
53  1  9 0.00 ==========================*
54  1  0 0.00 ===========================*
55  1  0 0.00 ===========================*
56  1 10 0.00 ==========================*
57  1  0 0.00 ========================*
58  1  0 0.00 =======================*
59  1 11 3.03 ===================* ----------------9                11
60  1  0 0.00 =================*
61  1 12 3.34 ==============* -------------------------1                9
62  1  0 0.00 =============*
63  1 13 3.56 ============* ------------------------------1            9
64  1  0 0.00 ===========*
65  1  0 0.00 ==========*
66  1  0 0.00 =========*
67  1 14 3.71 =========* ------------------------------------1*        9
68  1  0 0.00 =========*
69  1  0 0.00 ==========*
70  1  0 0.00 ============*
71  1 15 3.66 =============* ------------------------------1            9
72  1  0 0.00 =================*
73  1 16 3.42 ===================* ----------------------9             8
74  1  0 0.00 ======================*
75  1  0 0.00 =======================*
```

Figure 7: "...FOUR ONE..." (First Section)

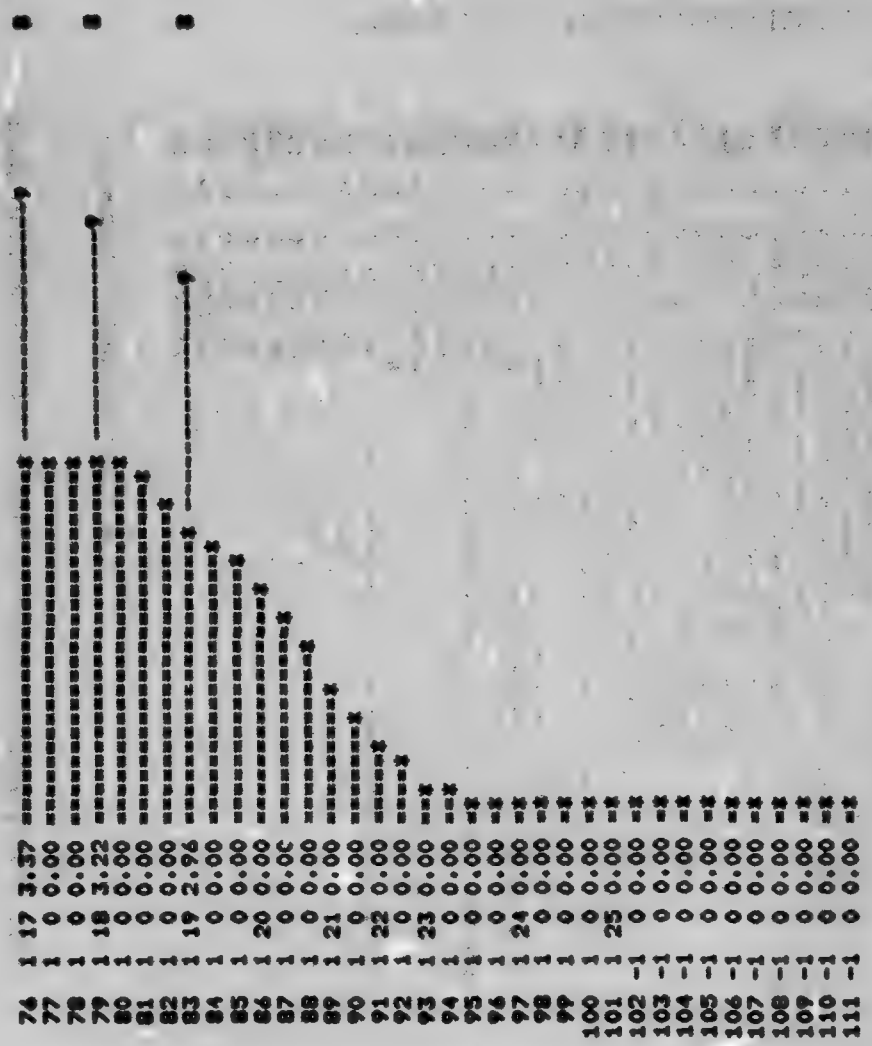


Figure 7a: "...FOUR ONE..." (Second Section)

```
TIME INTERVAL SAMPLES: FIRST= 25 LAST=113 NOISE LEVEL= 224.

THE WORDS RECOGNIZED ARE 3 8 8

25 -1  0 0.00 =*
24 -1  0 0.00 =*
27 -1  0 0.00 =*
28 -1  0 0.00 =*
29 -1  0 0.00 =*
30 -1  0 0.00 =*
31 -1  1 3.00 =*
32 -1  0 0.00 =*
33 -1  2 3.01 =*----------3                                                  11
34 -1  0 0.00 ==*
35 -1  0 0.00 ==*
36 -1  0 0.00 =*
37 -1  3 3.20 =*----------3*                                                 9
38 -1  0 0.00 =*
39  1  4 3.07 =*----------3                                                  9
40  1  5 2.99 =======*----------8                                            8
41  1  0 0.00 ==========*
42  1  4 2.97 ==========*----------8                                         8
43  1  0 0.00 ==========*
44  1  7 0.00 ==============*
45  1  0 0.00 ==============*
46  1  0 0.00 ==============*
47  1  0 0.00 ==============*
48  1  8 0.00 ==============*
49  1  0 0.00 ==============*
50  1  0 0.00 ==============*
51  1  0 0.00 ==============*
52  1  9 2.97 ==============*----------8                                      8
53  1  0 0.00 ==============*
54  1  0 0.00 ==============*
55  1  0 0.00 ==============*
56  1 10 3.31 ==============*----------8                                      8
57  1  0 0.00 ==============*
58  1 11 3.49 ==============*----------8*                                     8
59  1  0 0.00 ==============*
60  1  0 0.00 ==============*
61  1 12 3.28 ==============*----------8                                      8
62  1  0 0.00 ==============*
63  1  0 0.00 ==============*
64  1  0 0.00 ==============*
65  1 13 2.89 ==============*----------8                                      8
66  1  0 0.00 ==============*
67  1  0 0.00 ==============*
68  1  0 0.00 ==============*
69  1 14 0.00 ==============*
70  1  0 0.00 ==============*
71  1  0 0.00 ==============*
72  1 15 0.00 ==============*
73  1  0 0.00 ==============*
74  1 16 2.86 ==============*----------3                                      11
75  1  0 0.00 ==============*
76  1 17 3.31 ==============*----------3                                      11
77  1  0 0.00 ==============*
78 -1  0 0.00 ==============*
79 -1 18 3.50 ==============*----------3                                      9
```

Figure 8: "...THREE EIGHT EIGHT..." (First Section)

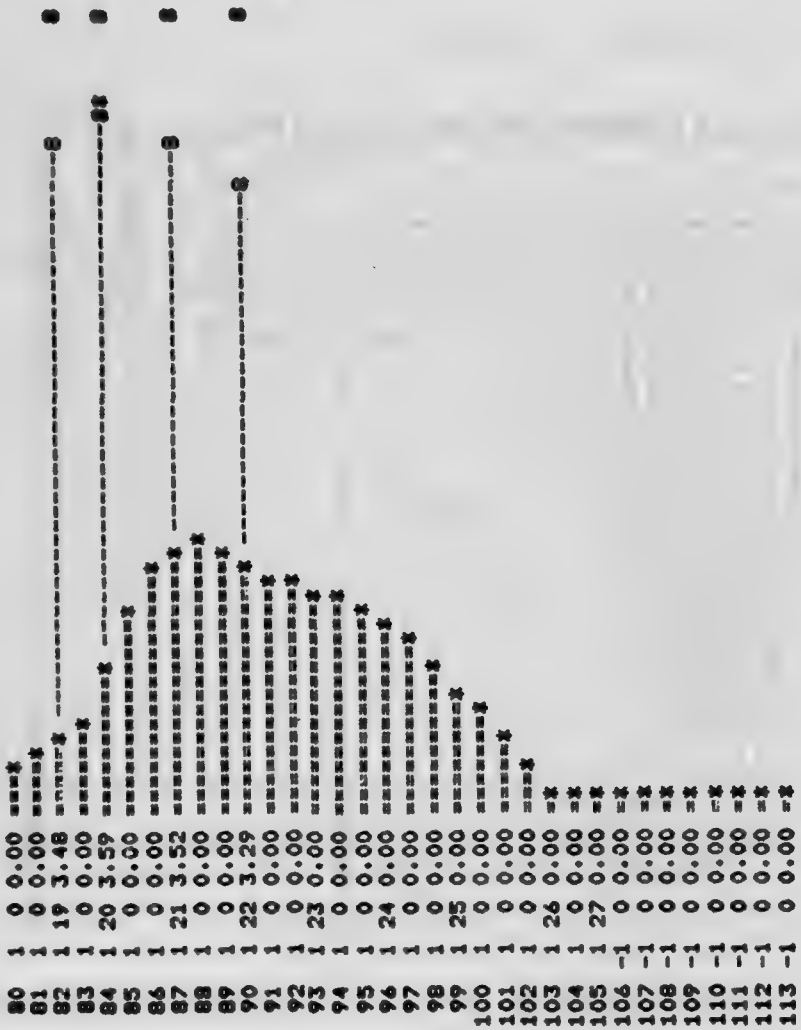


Figure 8a: "...THREE EIGHT EIGHT..." (Second Section)

```
TIME INTERVAL SAMPLES: FIRST= 20 LAST=176 NOISE LEVEL= 202.

THE WORDS RECOGNIZED ARE 1 2 3 4

20 -1  0 0.00 =*
21 -1  0 0.00 =*
22 -1  0 0.00 =*
23 -1  0 0.00 =*
24 -1  0 0.00 =*
25 -1  0 0.00 =*
26 -1  1 3.40 =* ----------------------------------------4          10
27 -1  0 0.00 =*
28 -1  2 3.29 =* ----------------------------------------1          10
29  1  0 0.00 =*
30  1  0 0.00 ==*
31  1  3 3.65 ===* ------------------------------------------1*      9
32  1  0 0.00 ====*
33  1  0 0.00 =====*
34  1  4 3.59 ========* ---------------------------------------1      9
35  1  0 0.00 ==========*
36  1  5 3.48 =============* ------------------------------9          8
37  1  0 0.00 ================*
38  1  6 3.30 =======================* -------------------9          8
39  1  0 0.00 ==========================*
40  1  7 2.91 ==============================* -------9                8
41  1  0 0.00 ============================*
42  1  8 0.00 ============================*
43  1  0 0.00 ============================*
44  1  0 0.00 ========================*
45  1  9 0.00 =====================*
46  1  0 0.00 ===============*
47  1 10 0.00 ==========*
48  1  0 0.00 =======*
49  1 11 0.00 ======*
50  1  0 0.00 =====*
51  1  0 0.00 ====*
52  1  0 0.00 ====*
53  1 12 0.00 ===*
54  1  0 0.00 ==*
55  1  0 0.00 =*
56  1 13 0.00 =*
57  1  0 0.00 =*
58  1 14 0.00 =*
59 -1 15 0.00 ==*
60 -1 16 3.03 ====* ------------------------------2               9
61 -1  0 0.00 =====*
62 -1  0 0.00 =====*
63 -1 17 3.47 =====* ------------------------------------2*        9
64 -1  0 0.00 =====*
65 -1 18 3.26 ====* ------------------------------------2           9
66 -1  0 0.00 ===*
67  1 19 2.84 =======* ---------------------------2                 9
68  1 20 2.88 ==========* ---------------------------8               8
69  1 21 2.95 ================* ---------------------8               9
70  1  0 0.00 =======================*
71  1 22 0.00 ==========================*
72  1  0 0.00 ============================*
73  1  0 0.00 ============================*
74  1  0 0.00 ============================*
```

Figure 9: "...ONE TWO THREE FOUR..." (First Section)

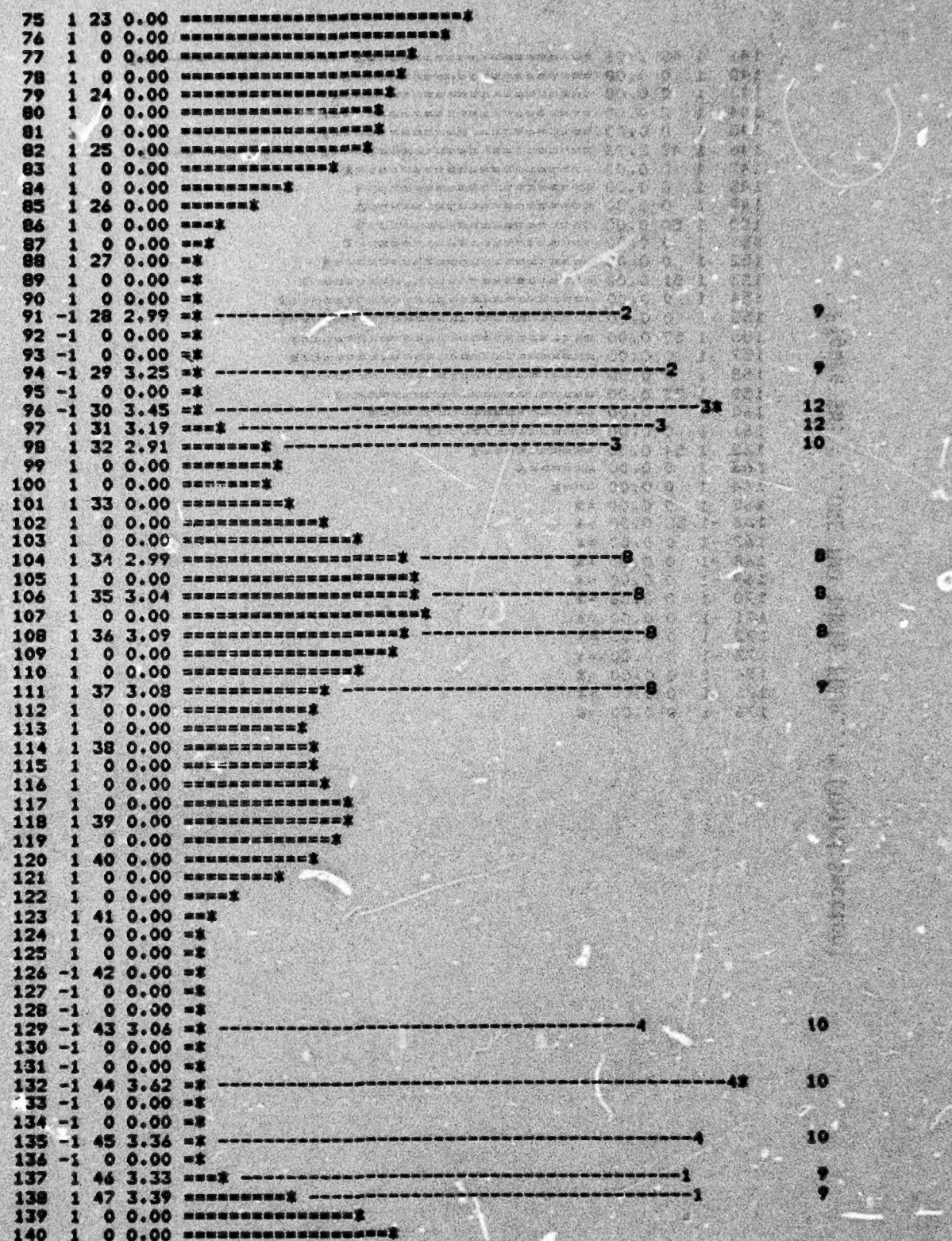


Figure 9a: "...ONE TWO THREE FOUR..." (Second Section)

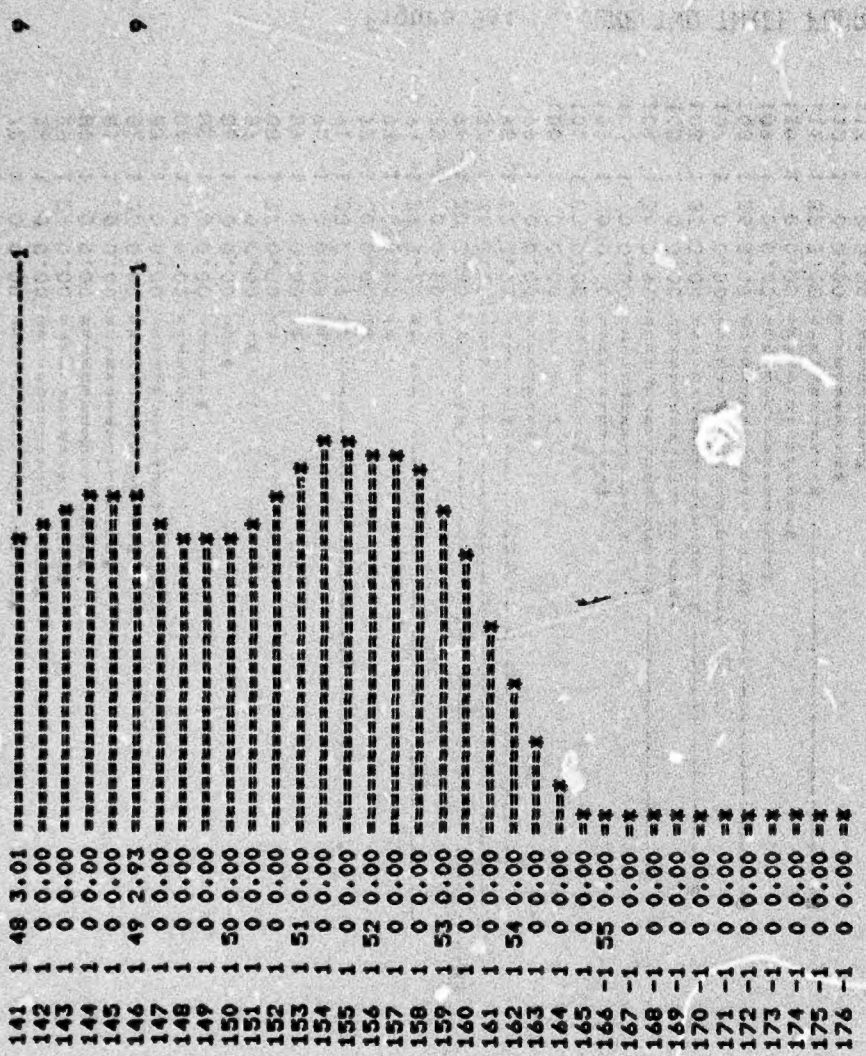
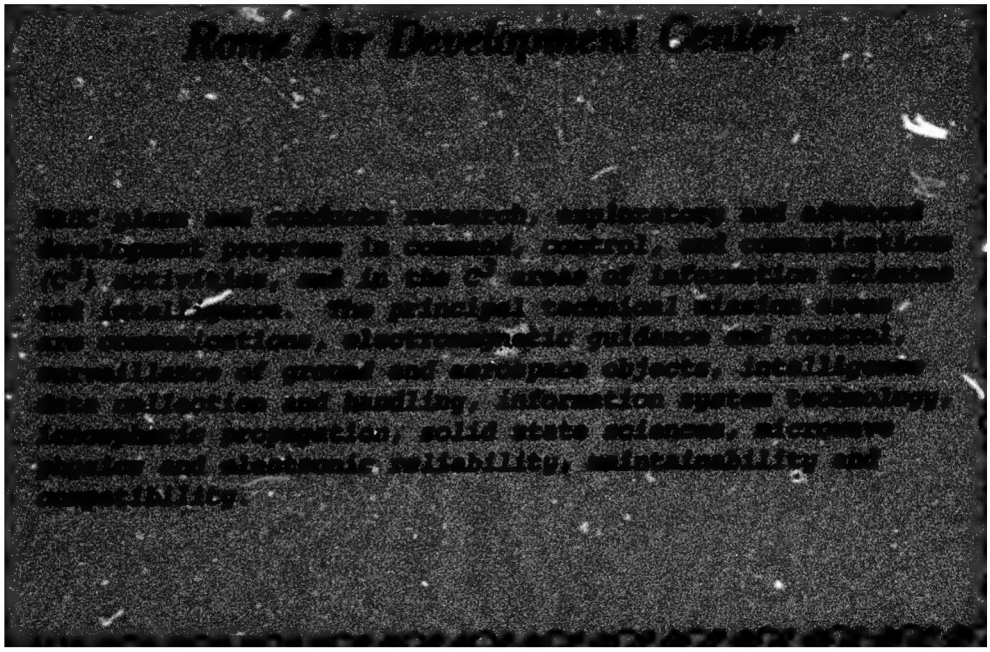


Figure 9b: "...ONE TWO THREE FOUR..." (Third Section)


Rome Air Development Center
RADC plans and conducts research, exploratory and advanced development programs in command, control, and communications (C3) activities, and in the C3 areas of information sciences and intelligence. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.